

# PATENT ABSTRACTS OF JAPAN

(11) Publication number : 2000-112946  
(43) Date of publication of application : 21.04.2000

(51) Int. Cl.

G06F 17/30

(21) Application number : 10-281817

(71) Applicant : NIPPON TELEGR & TELEPH CORP  
<NTT>

(22) Date of filing : 02.10.1998

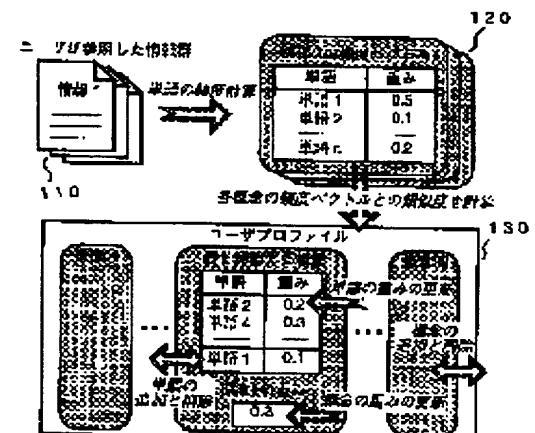
(72) Inventor : KIKKO HIROYUKI  
SATO NAOYUKI  
SUZUKI HIDEAKI  
SONEOKA AKINAO

## (54) INDIVIDUAL INFORMATION GENERATION METHOD AND INFORMATION SELECTION METHOD IN DISTRIBUTED RECOMMENDING SYSTEM AND RECORD MEDIUM RECORDING INDIVIDUAL INFORMATION GENERATION PROGRAM

### (57) Abstract:

PROBLEM TO BE SOLVED: To construct individual information which can be shared without the occurrence of contradiction in recommended information among plural recommending systems.

SOLUTION: Individual information is defined as the set of concept information. Concept information expresses individual concepts and it is constituted of the concept vector of the set containing the plural groups of [word and significance of word]. A frequency vector is set to be index information on the content of text information. The frequency vector of text information to which a user refers is calculated, similarity between the frequency vector and the concept vector of existed concept information is calculated, the most similar concept information is selected and concept information is updated according to whether similarity is smaller or larger than a reference value.



### LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

BEST AVAILABLE COPY

[Date of extinction of right]

Copyright (C) ; 1998, 2003 Japan Patent Office

(19)日本国特許庁 (JP)

## (12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-112946

(P2000-112946A)

(43)公開日 平成12年4月21日 (2000.4.21)

(51)Int.Cl.<sup>7</sup>

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/403

テーマコード(参考)

3 4 0 A 5 B 0 7 5

審査請求 未請求 請求項の数3 O.L (全6頁)

(21)出願番号 特願平10-281817

(22)出願日 平成10年10月2日 (1998.10.2)

(71)出願人 000004226

日本電信電話株式会社

東京都千代田区大手町二丁目3番1号

(72)発明者 橋高 博行

東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

(72)発明者 佐藤 直之

東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

(74)代理人 100073760

弁理士 鈴木 誠

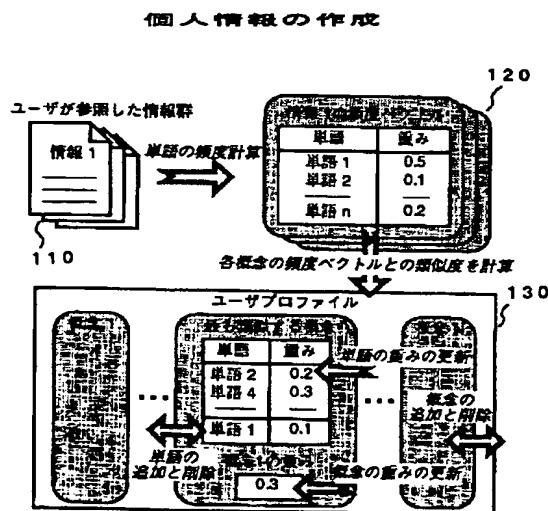
最終頁に続く

(54)【発明の名称】 分散リコメンドシステムにおける個人情報作成方法及び情報選択方法並びに個人情報作成プログラムを記録した記録媒体

(57)【要約】

【課題】 複数リコメンドシステム間で、推薦される情報に矛盾を生じさせずに共有できる個人情報を構築する。

【解決手段】 個人情報を概念情報の集合として定義する。概念情報は各々一つの概念を表現し、〔単語、単語の重要度〕の組を複数含む集合の概念ベクトルで構成する。頻度ベクトルはテキスト情報の内容のインデックス情報とする。利用者が参照したテキスト情報の頻度ベクトルを計算し、該頻度ベクトルと既存の各概念情報の概念ベクトルとの類似度を計算し、最も類似する概念情報を選択し、類似度が基準値より小さいか大きいかによって、概念情報を更新する。



1

## 【特許請求の範囲】

【請求項1】 ネットワークで接続された複数のリコマンドシステムにおいて、情報選択に利用される個人情報の作成方法であって、

個人情報を各々一つの概念を表現する概念情報の集合として定義し、各概念情報を〔単語、単語の重要度〕の組を複数含む概念ベクトルで構成し、テキスト情報の内容を示す〔単語、単語の重要度〕の組の集合を頻度ベクトルとし、

利用者が参照したテキスト情報の頻度ベクトルを計算し、

利用者の個人情報に含まれる概念情報の数が、既定値より少ない場合には、上記頻度ベクトルを概念ベクトルとする新しい概念情報を作成し、

概念情報の数が既定値より大きい場合は、既存の各概念情報の概念ベクトルと上記頻度ベクトルとの類似度を計算して最も類似する概念情報を選択し、類似度が基準値より小さい場合は、既存の概念情報の数を減らして、新たな概念情報として上記頻度ベクトルを持つ情報を追加し、類似度が基準値より大きい場合は、上記頻度ベクトルを、選択した概念情報の持つ頻度ベクトルに加え合わせることを特徴とする分散リコマンドシステムにおける個人情報作成方法。

【請求項2】 請求項1で作成された個人情報を用いた情報の選択方法であって、テキスト情報の各々の頻度ベクトルと、個人情報の各概念情報の概念ベクトルとの類似度を計算し、概念情報ごとに、類似度の高いテキスト情報を必要数だけ選択することを特徴とする分散リコマンドシステムにおける情報選択方法。

【請求項3】 ネットワークで接続された複数のリコマンドシステムにおいて情報選択に利用される個人情報作成のためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

利用者が参照したテキスト情報の頻度ベクトルを計算する処理プロセスと、

利用者の個人情報に含まれる概念情報の数が、既定値より少ない場合には、上記頻度ベクトルを概念ベクトルとする新しい概念情報を作成する処理プロセスと、 概念情報の数が既定値より大きい場合は、既存の各概念情報の概念ベクトルと上記頻度ベクトルとの類似度を計算して最も類似する概念情報を選択し、類似度が基準値より小さい場合は、既存の概念情報の数を減らして、新たな概念情報として上記頻度ベクトルを持つ情報を追加し、類似度が基準値より大きい場合は、上記頻度ベクトルを、選択した概念情報の持つ頻度ベクトルに加え合わせることを特徴とする個人情報作成プログラムを記録した記録媒体。

## 【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ネットワーク上で

2

利用されるリコマンドシステムにおける、個人情報の作成方法及びこの個人情報を用いた情報選択方法並びに個人情報作成プログラムを記録した記録媒体に関する。

【0002】

【従来の技術】現在、インターネットなどの広域ネットワークにおいては、利用者の求める情報をあらかじめ予測し、これを推薦するリコマンドシステムが多数提案されている。しかし、これらのシステムは、個々別々に開発されており、各々がそれぞれ別々に利用者の情報を収集する処理を行っている。このため、複数の異なったシステムによって推薦される情報には、その内容に矛盾が起こる可能性は高く、全く関連のない情報がそれぞれ推薦されることもある。

【0003】

【発明が解決しようとする課題】これまで、複数システム間での個人情報の統一が困難であった一因には、作成された情報の解釈を統一することが難しいという問題があった。例えば個人情報の作成手法として、単語とその出現度数を参考にする手法は数多く提案されている。これららの手法では、個々の単語に注目し、利用者が参照したテキスト情報に使用された単語から個人情報を作成する。このようにして作成された個人情報は、単語をキーワードとして検索を行うようなシステムには非常に適している。しかし、情報を単語レベルではなく、さらに抽象化した「概念」として利用するようなシステムにおいては、この情報は役立たない。また、通常単語には複数の異なった意味があるので、この形式の個人情報は人（管理者等）が見て理解することが難しいという欠点もある。

【0004】また別の作成手法として、システムが定めた「概念」の基準に対する相対的な値として個人情報を作成する手法も多々存在する。この場合は、異なったシステム間で同一「概念」を明確に定義することが難しく、複数システムで推薦する情報に矛盾が生じないようすることは極めて難しい。

【0005】本発明の目的は、上述のような問題を解決し、ネットワークで接続された複数のシステムにおいて同時に利用できる個人情報の作成方法及びこの情報を用いた標準的な情報選択方法を提供し、更には個人情報作成プログラムを記録した記録媒体を提供することにある。

【0006】

【課題を解決するための手段】本発明では、個人情報（ユーザプロファイル）を「単語とその重要度の集合」と、それによって定義される「概念」で構築する。「概念」を、より明確な「単語とその重要度の集合」によって示すことで、個人情報のリコマンドシステム間での相互利用を容易にし、推薦される情報間に矛盾を生じさせなくすることが可能になる。

【0007】

【発明の実施の形態】初めに、実施の形態で用いる用語を説明する。

#### (1) 概念情報

一つの概念情報は一つの概念を表現するためにある。本発明では、個人情報を概念情報の集合として定義し、各利用者の個人情報は、概念情報を各々最大既定値Nだけ保持することができるとする。初期化段階においては個人情報は概念情報を持っていない。利用者がその嗜好によって、テキストなどの情報を参照する度に、自動的に構築・更新される。

【0008】概念情報は以下の要素を含む。

概念ベクトル：頻度ベクトルは、概念を表現し、多システム間で計算機によって取り扱えるようにするためにある。概念ベクトルは〔単語、単語の重要度〕の組を最大M個含む集合であるとする。このよう、〔単語、単語の重要度〕を組として保存する手法は、従来から数多く提案されている。特に、後述のTFIDF (Term Frequency and Inverse Document Frequency) と呼ばれる方法は、テキスト情報に対して、そのインデックス情報として上記の集合（ただしMは固定値ではない）を計算する。この手法は非常に有名かつ有効である。  
\* 20

$$\text{Sim}(X, Y) = \frac{X \text{ と } Y \text{ の内積}}{|X| \times |Y|}$$

この計算では、各々のベクトルの向く方向の違いに注目しており、方向が近付くほど類似度Sim(X, Y)の値は大きくなる。

#### 【0011】(4) TFIDF

テキストから、その文中に出現する全ての単語（名詞のみの場合もある）を抽出し、各単語の出現頻度を利用してテキスト全体を示す情報として用いる手法の一つである。TFIDFでは、次の処理によって対象としたテキストの内容を要約した特徴量である多次元多値ベクトルを生成する。なお、特徴情報を計算する全てのテキストは集められているとする。

【0012】特徴情報を計算したいテキストXについて、そのテキストで使用される全単語（名詞のみの場合もある）と、その単語の当該テキスト中で出現度数を調べる。TF (Term Frequency) を以下のように定める。

TF = 単語kがテキストXで使用された回数

テキストX中で使用された全ての単語について、これらが各々全テキスト中で何個のテキストに出現するか調べる。この情報を用いてIDF (Inverse Document Frequency) を以下のように定める。

IDF(k) = f(n(k), N)

n(k) = 単語(k) = 単語kが（全テキスト中で）使用されているテキストの数

N = 全テキストの数

f(x, y) は任意の関数

f(x, y) = log(y/x) であることが多い

\* 重要度：概念情報間における、重要性の度合を表現する。この値が高い程、利用者は当該概念に対して深い興味を持つことを意味する。

#### 【0009】(2) 頻度ベクトル

頻度ベクトルは、テキスト情報の内容を示すためのインデックス情報である。概念ベクトルと同様に、〔単語、単語の重要度〕の組の集合で表現する。概念情報がN個未満の場合、参照した情報の頻度ベクトルが一つの概念情報となる。この情報は既存のTFIDF等の技術を用いて計算する。

#### 【0010】(3) 類似度の計算

頻度ベクトルと概念ベクトル、あるいは概念ベクトル同士がどのくらい関連を持っているかの度合を類似度と呼んでいる。類似度が高いほど、これらは関連が深い。類似度の計算は、〔単語、単語の重要度〕の集合を、各単語が軸を示し、その重要度が値を示していると考えた多次元多値ベクトルを用いて行う。二つのベクトルをX及びYとすると、類似度Sim(X, Y)の計算は以下のように行う。

#### 【数1】

$$(0 \leq \text{Sim}(X, Y) \leq 1)$$

テキストXで使用される全ての単語について、以下の値を計算する。

$$\text{TFIDF}(X, k) = \text{TF}(X, k) \times \text{IDF}(k)$$

【0013】各単語ごとにTFIDF値を求める、単語を各々軸とした多次元多値ベクトルとなる。本実施の形態では、この多次元多値ベクトルを概念ベクトルや頻度ベクトルとして扱い、単語の重要度にTFIDF値を使用する。

【0014】次に、本発明の一実施の形態の処理について詳述する。処理は、概念情報の集合である個人情報を作成する処理と、この個人情報を用いて情報を選択する処理に大別される。

【0015】〔個人情報の作成〕図1は個人情報作成の概念図を示したものである。110はユーザが参照した情報群、120は情報の頻度ベクトル、130は個人情報のユーザプロファイルである。初期化段階においては、個人情報（プロファイル）は概念情報を持たない。

【0016】図2に、個人情報作成の処理フローを示す。この処理は、ユーザが情報を参照する度に繰り返される。

(1) ユーザが参照したテキスト情報110から単語を抽出する。テキスト情報が日本語であった場合には、形態素解析等の手段を利用する必要がある。抽出された各単語に対して、その単語のテキスト中での出現度数を調べ、頻度ベクトル120を計算する。頻度ベクトルの計算には、既存のTFIDF等を利用する。

50 (2) ユーザの個人情報（ユーザプロファイル）130

に含まれる概念情報の数が、既定値Nより少ない場合には、上記(1)の頻度ベクトル120を概念ベクトルとする新しい概念情報を作成し、該概念情報の重要度を例えば「1」とする。この際、頻度ベクトルの要素[単語、単語の重要度]の数が、既定値Mよりも大きかった場合には、単語の重要度の大きい順にM個を選択し、これを概念ベクトルとする。この場合、この時点で1図の作成処理が終了する。

(3) 概念情報の数が既定値Nより大きかった場合は、既存の各概念情報の概念ベクトルと、上記(1)の頻度ベクトルとの類似度を計算し、最も類似する概念情報を選択する。そして、この選択された概念情報との類似度によって、次の(i)、(ii)のどちらかの処理を行う。

【0017】(i) 類似度が基準値より小さかった場合

既存の概念情報の数を一つ減らしN-1とする。この方法として、既存の概念情報のうち最も重みの少ないものを削除する方法と、類似する概念情報同士を一つに合成する方法の二つがある。なお、最も重みの少ないものを削除する方法において、その候補が複数存在する場合には、最も古く作成された概念情報を削除する。概念情報を削減した後、新たな概念情報として、上記(1)の頻度ベクトル120を持つ情報を追加する。追加時に[単語、単語の重要度]の個数が既定値Mより大きかった場合は、上記(2)と同様の処理を行って、個数をMとする。

【0018】(ii) 類似度が基準値より大きかった場合

上記(1)の頻度ベクトルを、選択した概念情報の持つ頻度ベクトルに加え合わせる。具体的には、頻度ベクトルの各[単語、単語の重要度]に対して、概念ベクトルに同一の単語を持つ組があった場合、この組の単語の重要度を頻度ベクトルの組の単語の重要度分だけ増加させる。また、概念ベクトルに同一の単語を持つ組がなかった場合、新たに[単語、単語の重要度]の組を作成し、これを概念ベクトルの集合に加える。この加え合わせの処理によって、[単語、単語の重要度]の個数が既定値Mより大きくなる場合は、(2)と同様に単語の重要度の高いM個を選択して残りを削除する。また、同時に、概念情報の重要度を既定値Sだけ増加させる。

【0019】〔情報の選択〕図3は、情報選択の概念図を示したものである。310は上記のようにして作成された利用者の個人情報(ユーザプロファイル)、320や330は選択(推薦)される情報群である。

【0020】図4に情報選択の処理フローを示す。この処理は、各リコマンドシステムにおいて、利用者から情報要求がある毎に繰り返される。リコマンドシステムでは、複数のテキスト情報の集合から、利用者へ推薦すべき情報を次のようにして選択する。

(1) 各々のテキスト情報の頻度ベクトルと、ユーザプロファイル310の各概念情報の概念ベクトルとの類似度を計算する。

(2) 概念情報ごとに、類似度の高いテキスト情報を必要数だけ選択する。

(3) 選択した情報を、該当概念に属する情報320として利用者に提示する。この際、概念情報の持つ重要度を利用して、重要度の高い順に情報330を表示することもできる。

【0021】図5に、本発明を用いた分散リコマンドシステムの第1の実現例を示す。これは、利用者側(利用者端末)において個人情報を直接管理する例である。利用者aは、例えばリコマンドシステムAに情報を要求する際、一緒に個人情報を送付する(①)。リコマンドシステムAは、送付された個人情報を使用して情報の選択処理を実施し、推薦情報を利用者aへ送付する(②)。利用者aでは、該送付された情報を利用して個人情報(ユーザプロファイル)を更新する(③)。これにより、ユーザの好みの変化に応じて、個人情報は常に最新のものが構築される。

【0022】図6に、本発明を用いた分散リコマンドシステムの第2の実現例を示す。これは、ネットワーク上に、個人情報を一括して管理するサーバ(個人情報管理データベース)600を設ける例である。個人情報管理データベース600は、利用者IDを見出しとして各利用者の個人情報を管理する。利用者aは、リコマンドシステムAに情報を要求する(①)。リコマンドシステムAは、利用者aの個人情報を個人情報管理データベース600に要求し(②)、個人情報管理データベース600は該当個人情報をリコマンドシステムAに送付する

(③)。リコマンドシステムAは、該個人情報を使用して情報の選択処理を実施し、推薦情報を利用者aへ送付する(④)。同時に、リコマンドシステムAは、該利用者aに送付した情報を個人情報管理データベース600へも送付する(⑤)。個人情報管理データベース600では、該送付された情報を利用して利用者aの個人情報(ユーザプロファイル)を更新する(⑥)。

【0023】以上、本発明の一実施の形態を説明したが、図3に示したような処理フローチャートをコンピュータに実行させるためのプログラムは、該コンピュータが読み取り可能な記録媒体、例えばフロッピーディスクやメモリカード、コンパクトディスク(CD-ROM)などに記録して提供することが可能である。

【0024】  
【発明の効果】以上説明したように、本発明では、個人情報を「単語とその重要度」と、それによって定義される「概念」で構築することによってネットワークで接続された複数のリコマンドシステムにおいて、推薦される情報間に矛盾を生じさせないようにすることが可能になる。

## 【画面の簡単な説明】

【図1】本発明による個人情報の作成の概念図である。

【図2】個人情報作成の処理フロー図である。

【図3】本発明による個人情報を利用したテキスト情報選択の概念図である。

【図4】テキスト情報選択の処理フロー図である。

【図5】本発明による分散リコメンドシステムの第1の実現例である。

\* 【図6】本発明による分散リコメンドシステムの第2の実現例である。

## 【符号の説明】

110 ユーザ参照情報群

120 情報の頻度ベクトル

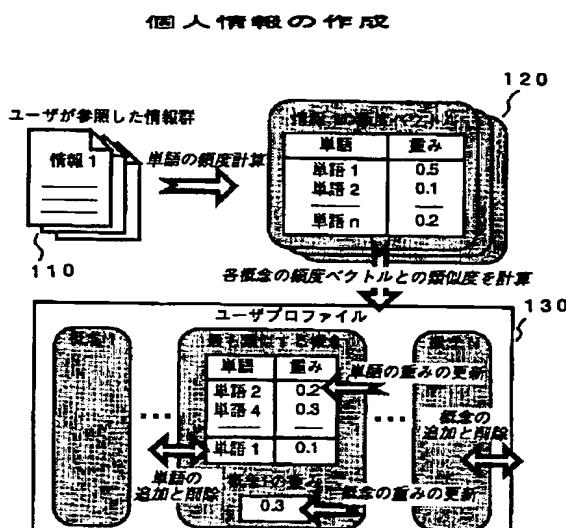
130 作成された個人情報

310 選択に利用される個人情報

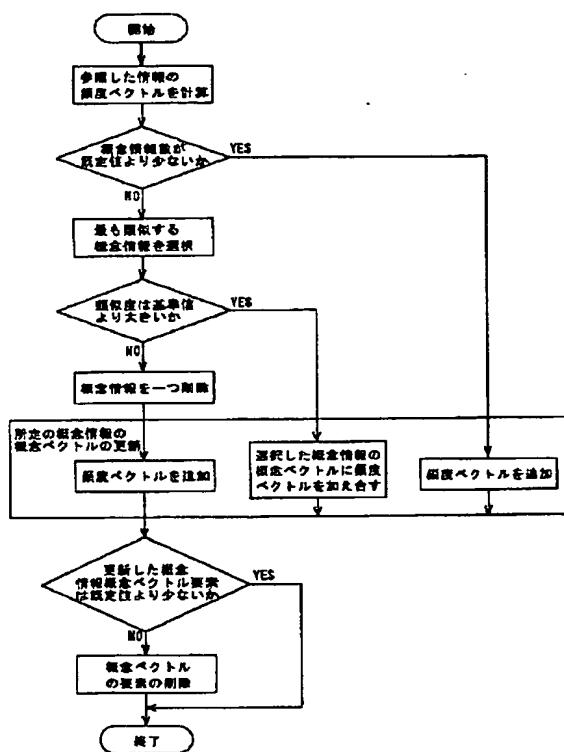
320, 330 推奨情報群

\*

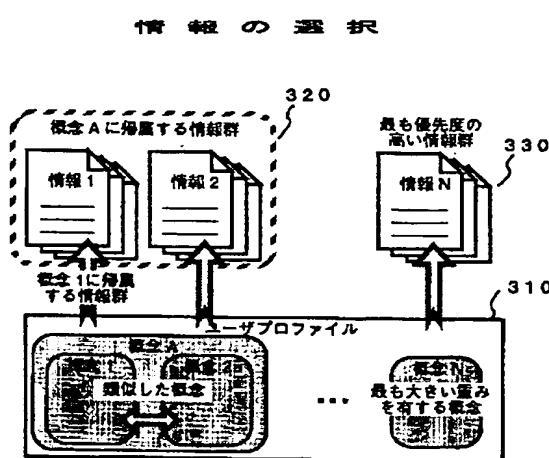
【図1】



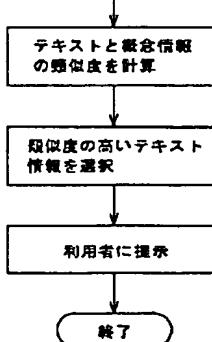
【図2】



【図3】

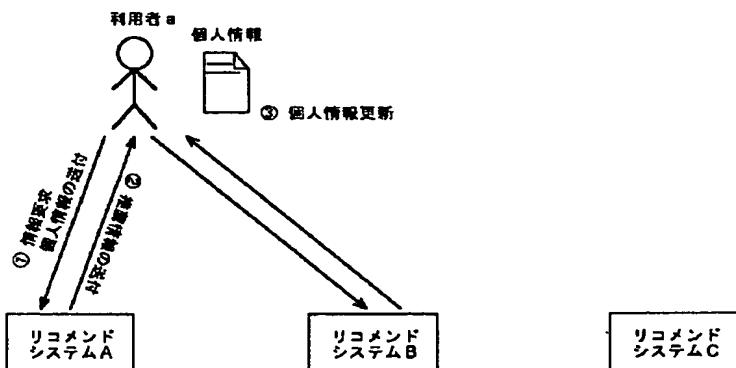


【図4】

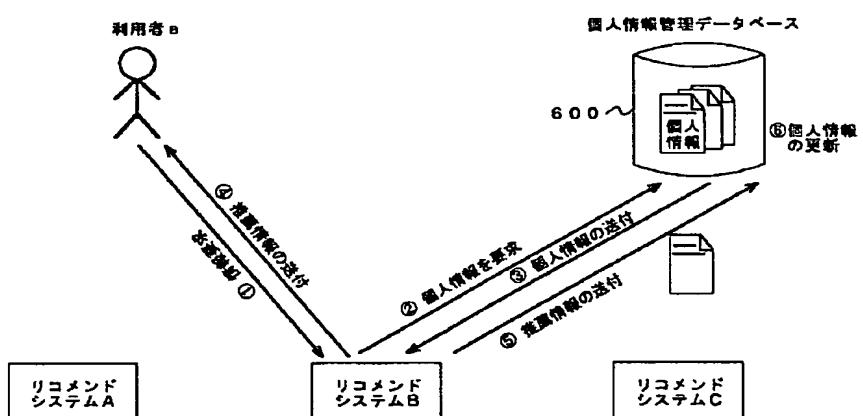


BEST AVAILABLE COPY

【図5】



【図6】



フロントページの続き

(72)発明者 鈴木 英明  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

(72)発明者 曽根岡 昭直  
東京都新宿区西新宿三丁目19番2号 日本  
電信電話株式会社内

F ターム(参考) 5B075 KK07 KK40 ND03 NK02 NK32  
NR10 NR20 PP28 PR06 PR08  
QM08 UU06 UU40